BIONUMERICS Tutorial:

wgSNP: analysis and clustering

1 Introduction

A Single Nucleotide Polymorphism (SNP) is a variation in a single nucleotide, which occurs at a specific position of the genome. When performed on whole genome sequences (WGS), this analysis is referred as **whole genome SNP (wgSNP) analysis**. When performed in BIONUMERICS a typical workflow for wgSNP analysis consists of following steps:

1. Choose a reference sequence
2. Map sequence reads against the reference sequence (locally or on the cloud calculation engine)
3. Perform wgSNP analysis and filter out relevant SNPs
4. wgSNP clustering

In this tutorial we will focus on the last two steps of the workflow. The first two steps are covered in the tutorials ”wgSNP with mapping performed on the cloud calculation engine” and ”wgSNP with mapping performed locally on your own computer”.

2 Preparing the database

The **WGS demo database** for *Staphylococcus aureus* contains entries for which a reference mapping has already been calculated. The resulting sequences are stored in the **SNP outbreak** sequence type and will be used in this tutorial to illustrate the *SNP filtering* window, the use of SNP filters, and possible follow-up analysis tools available in BIONUMERICS.

This demo database can be downloaded directly from the *BIONUMERICS Startup* window (see 2.1), or restored from the back-up file available on our website (see 2.2).

2.1 Option 1: Download demo database from the Startup Screen

1. To download the database directly from the *BIONUMERICS Startup* window, click the  button, located in the toolbar in the *BIONUMERICS Startup* window.

This calls the *Tutorial databases* window (see Figure 1).

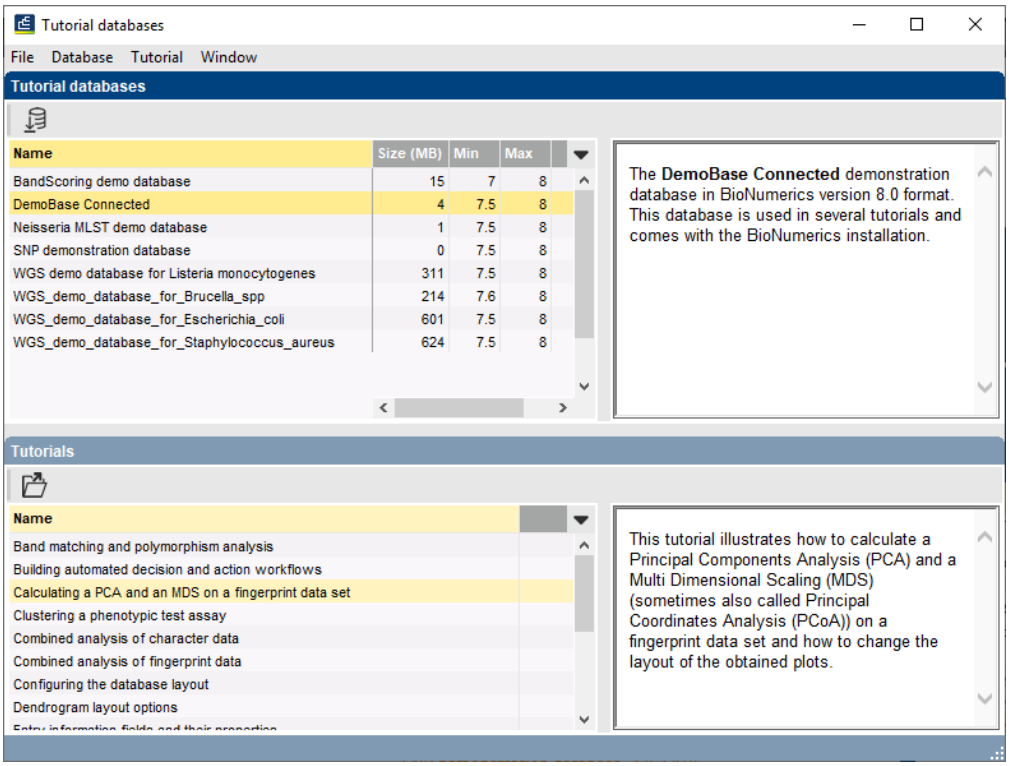

Figure 1: The *Tutorial databases* window, used to download the demonstration database.

2. Select the **WGS_demo_database_for_Staphylococcus_aureus** from the list and select ***Database*** > ***Download*** ( ).

3. Confirm the installation of the database and press <***OK***> after successful installation of the database.

4. Close the *Tutorial databases* window with ***File*** > ***Exit***.

The **WGS_demo_database_for_Staphylococcus_aureus** appears in the *BIONUMERICS Startup* window.

5. Double-click the **WGS_demo_database_for_Staphylococcus_aureus** in the *BIONUMERICS Startup* window to open the database.

2.2 Option 2: Restore demo database from back-up file

A BIONUMERICS back-up file of the whole genome demo database for Staphylococcus aureus is also available on our website. This backup can be restored to a functional database in BIONUMERICS.

6. Download the file `wgMLST_SAUR.bnbk` file from https://www.applied-maths.com/download/sample-data, under 'WGS_demo_database_for_Staphylococcus_aureus'.

In contrast to other browsers, some versions of Internet Explorer rename the `wgMLST_SAUR.bnbk` database backup file into `wgMLST_SAUR.zip`. If this happens, you should manually remove the `.zip` file extension and replace with `.bnbk`. A warning will appear ("If you change a file name extension, the file might become unusable."), but you can safely confirm this action. Keep in mind that Windows might not display the `.zip` file extension if the option "Hide extensions for known file types" is checked in your Windows folder options.

7. In the *BIONUMERICS Startup* window, press the button. From the menu that appears, select ***Restore database...***.
8. Browse for the downloaded file and select ***Create copy***. Note that, if ***Overwrite*** is selected, an existing database will be overwritten.
9. Specify a new name for this demonstration database, e.g. "Whole genome Staphylococcus aureus demobase".
10. Click <***OK***> to start restoring the database from the backup file (see Figure 2).

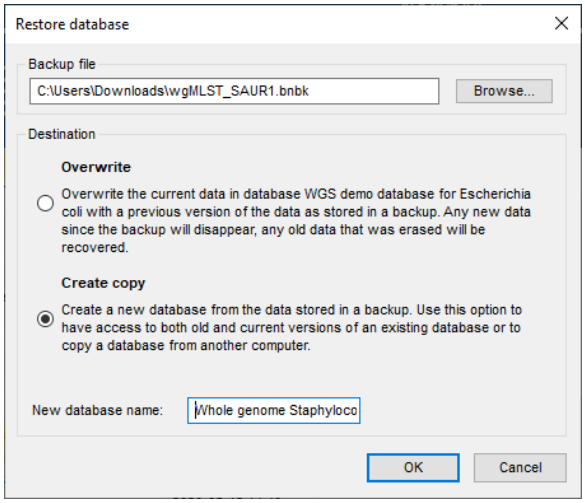

Figure 2: Restoring the whole genome demonstration database from the BioNumerics backup file `wgMLST_SAUR.bnbk`.

11. Once the process is complete, click <***Yes***> to open the database.

The *Main* window is displayed (see Figure 3).

3 Perform wgSNP analysis and filter out revelant SNPs

In our demonstration database, a reference mapping has already been calculated for the entries from the **Neonatal MRSA study**. The resulting sequences are stored in the **SNP outbreak** sequence type and will be used in this section to start a SNP analysis and to illustrate the use of SNP filters in the *SNP filtering* window.

1. Make sure no selection is present in the *Database entries* panel: press the **F4**-key to clear any entry selection.
2. In the *Database entries* panel, select the **Neonatal MRSA study** view from the list (see Figure 4) and use ***Edit*** > ***Select all*** (**Ctrl+A**) to select all 14 entries contained in this study.

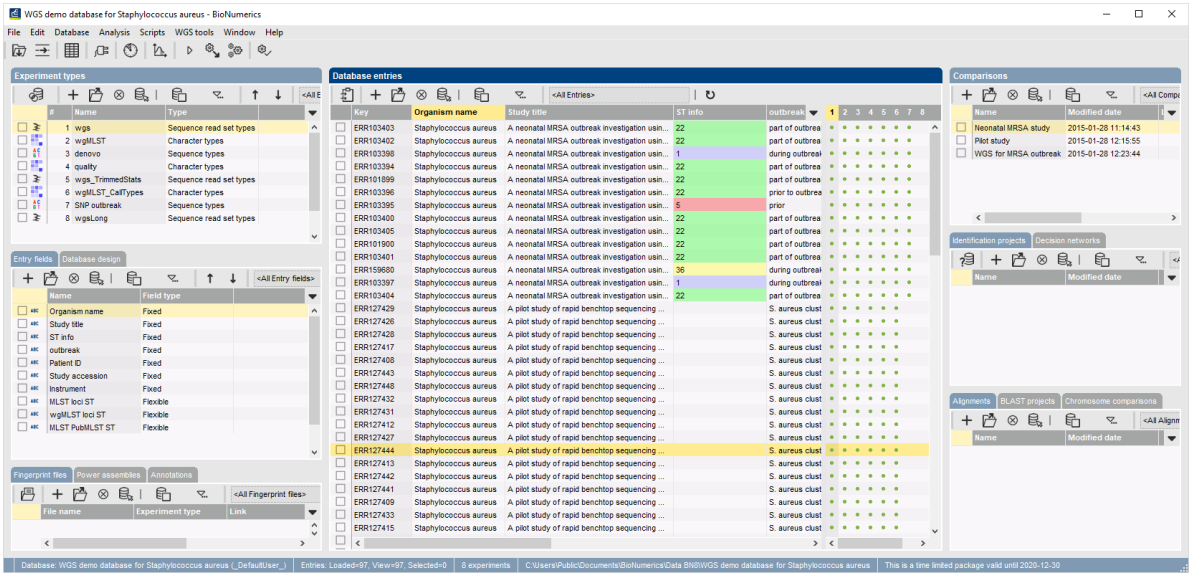

Figure 3: The *Staphylococcus aureus* demonstration database: the *Main* window.

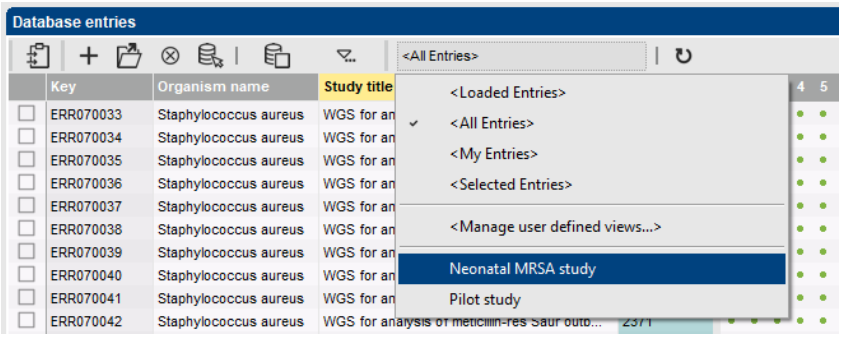

Figure 4: Select a view from the list.

3. Select ***Analysis*** > ***Sequence types*** > ***Start SNP analysis...*** to start the *SNP analysis* wizard.

4. Select **SNP outbreak** as ***Experiment type*** and press <***Next***> (see Figure 5).

A number of predefined SNP templates are available.

5. Highlight the ***Strict filtering*** template and press <***Next***> (see Figure 6).

6. Check ***Open SNP analysis window*** and press <***Finish***> (see Figure 7).

It will take a few moments to load the sequences and apply the filters from the SNP template. The resulting *SNP filtering* window is shown in Figure 8.

This window consists of following panels:

- The *Entries* panel shows all entries that are included in the SNP analysis, with all entry information fields. Two additional fields are present: 'Total' shows the raw number of SNPs (i.e. without any SNP filter applied) and 'Retained' shows the number of SNPs after applying all active SNP filters for the sample sequence.

- The *Filters* panel shows the list of SNP filters that are applied, with the 'Info' column showing additional information regarding the filter and applied settings (if applicable). This list is

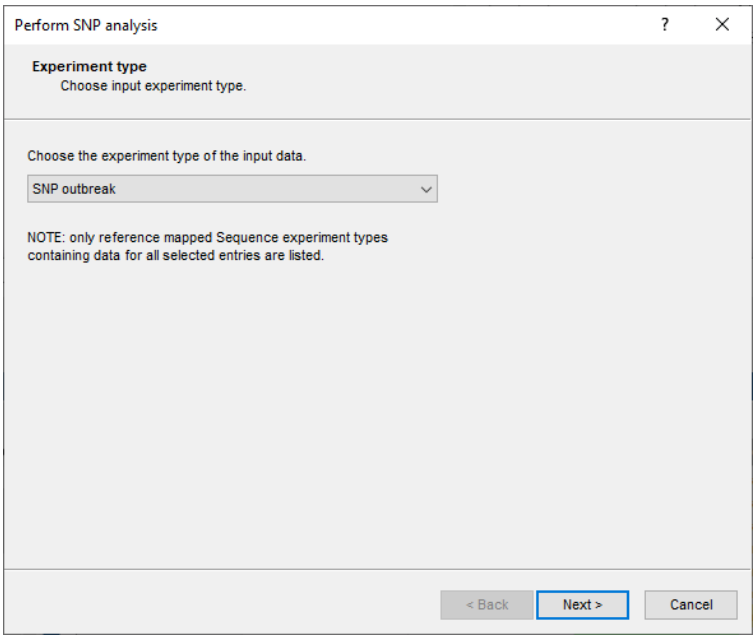

Figure 5: Select the input sequences.

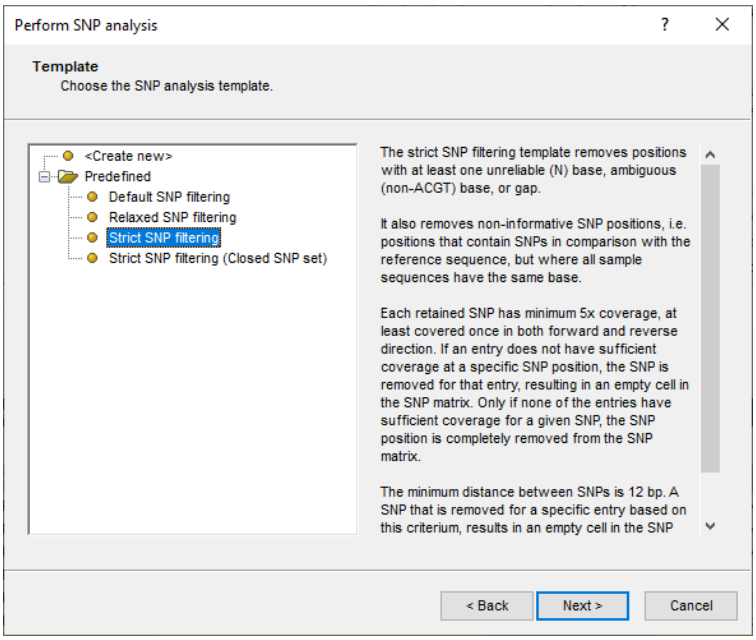

Figure 6: Choose a SNP filtering.

initially populated from the SNP template, but SNP filters can be added or removed and their settings can be changed.

- The *SNP Positions* panel shows information on all positions where at least one SNP was detected. For each SNP filter that is listed in the *Filters* panel, a column is displayed with the filter's result on each position. The bottom of this panel shows a sub-panel with the details on the highlighted position, i.e. showing the base and filter results for all the sample sequences on that position.
- The *Entry SNPs* panel lists the SNPs for the highlighted entry in the *Entries* panel.
- The *Genome* panel shows the SNPs on a genome view.

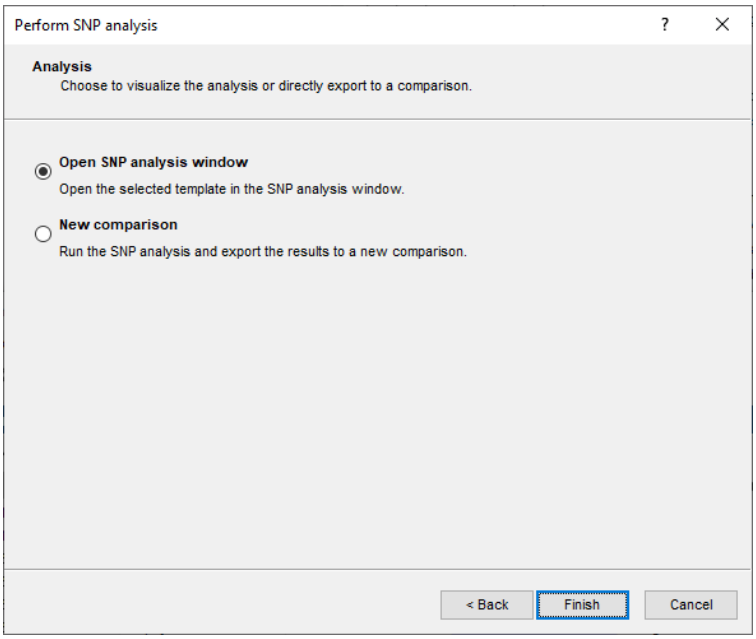

Figure 7: Open SNP analysis window.

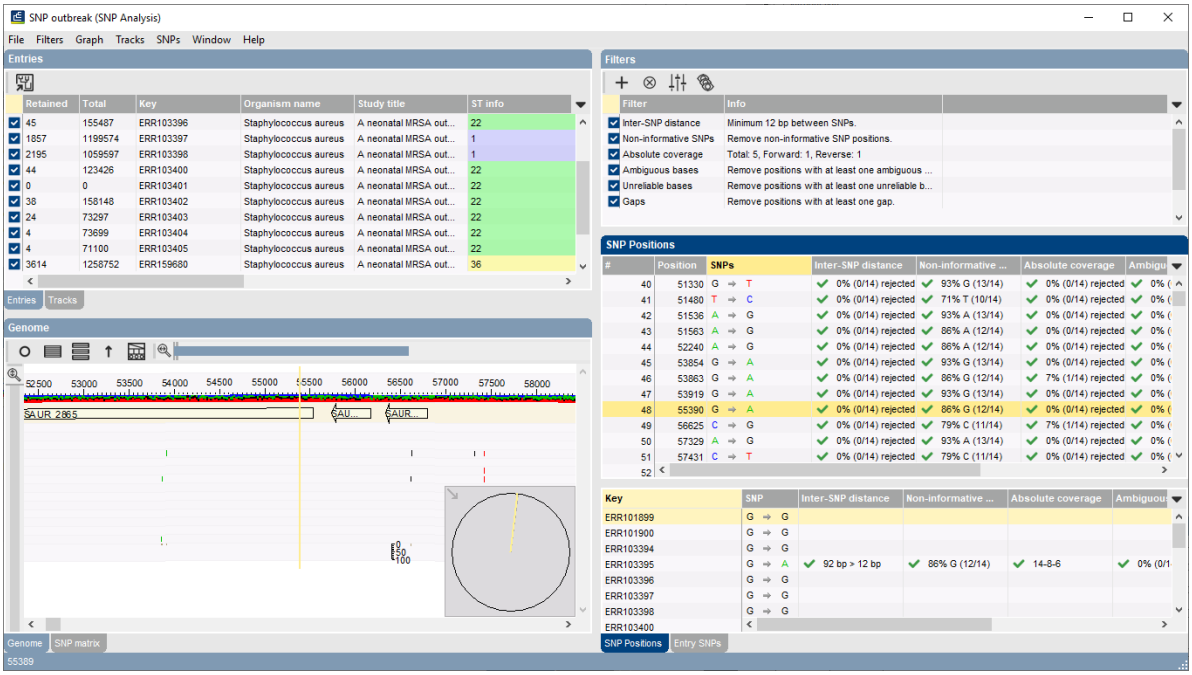

Figure 8: SNP analysis window.

- The *Tracks* panel in default view is displayed as a tab with the *Entries* panel. With this panel, you can determine which tracks are plotted in the *Genome* panel.
- The *SNP matrix* panel shows the resulting SNP matrix, as it would be exported.

Whenever possible, the cursor position is synchronized between the different panels:

7. Click on a position in the *SNP Positions* panel for example.

The details in the bottom part of the panel are updated and so is the *Genome* panel: the graph will show the position. Furthermore, the clicked position in the *SNP Positions* panel will appear

highlighted in the *Entry SNPs* panel, *only* if the currently highlighted entry in the *Entries* panel has a SNP at that position.

8. Double-click a position in the details panel (bottom part of the *SNP Positions* panel) or in the *Entry SNPs* panel.

This action will open the *Sequence editor* window of the corresponding sequence, with this position highlighted. If a sequence assembly is available in BIONUMERICS, the will be active and selecting ***File*** > ***Open assembler*** ( ) will open the assembly.

Sequence assemblies are not available when the remapping was performed on the calculation engine.

9. A SNP filter can be added with ***Filters*** > ***Add filter...*** (+).

10. Check or uncheck an individual SNP filter in the *Filters* panel to view its effect.

When the toggle ***Filters*** > ***Toggle rejected SNP visibility*** is unchecked ( ), the positions in the *SNP Positions* panel and the *Entry SNPs* panel will be limited to the retained SNPs, i.e. those SNPs that have passed the applied SNP filters.

When the toggle is checked ( ) the listed positions in both panels correspond to the total (i.e., unfiltered) SNP set.

11. Click on the tab of the *SNP matrix* panel to show the SNP matrix (see Figure 9).

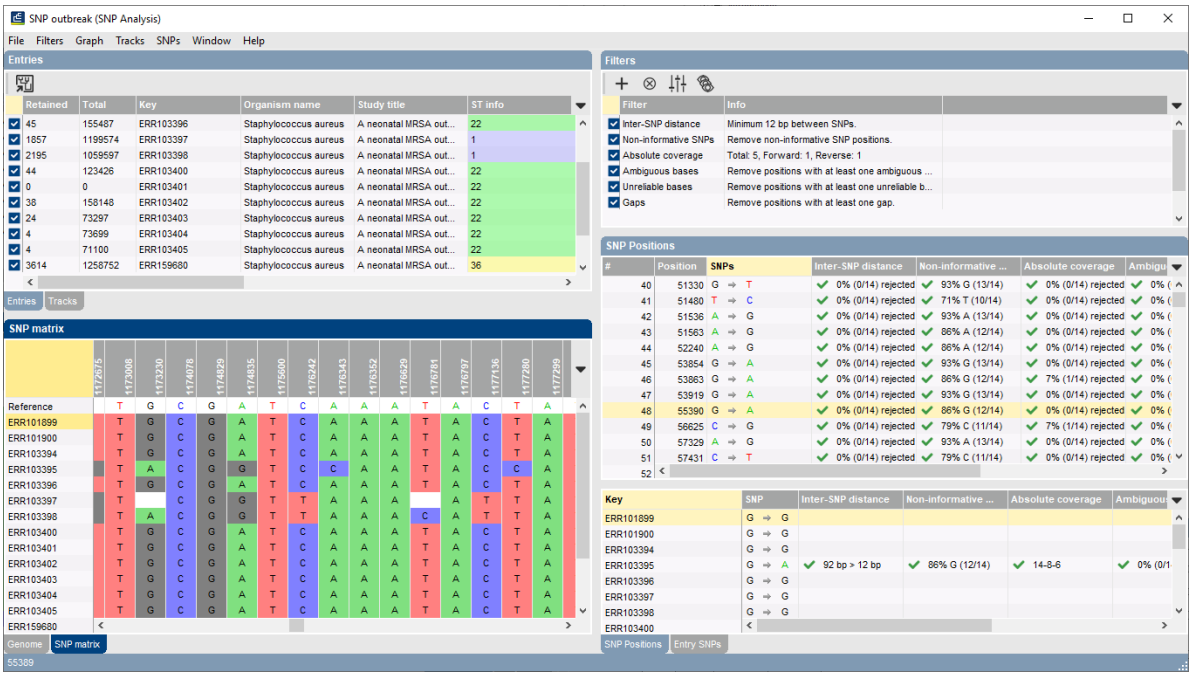

Figure 9: SNP matrix displayed.

12. Select ***File*** > ***Export to comparison...*** ( ) to export the SNP matrix to a comparison.

In the *Comparison* window a cluster analysis can be calculated based on the exported SNP data (see 4.1).

4 Follow-up analyis

4.1 Cluster analysis on SNP data

1. Selecting ***File*** > ***Export to comparison...*** () in the *SNP filtering* window exports the SNP matrix to a new comparison.

In this comparison, the SNP matrix is available as a *character aspect* of the **SNP outbreak** sequence experiment type (see Figure 10).

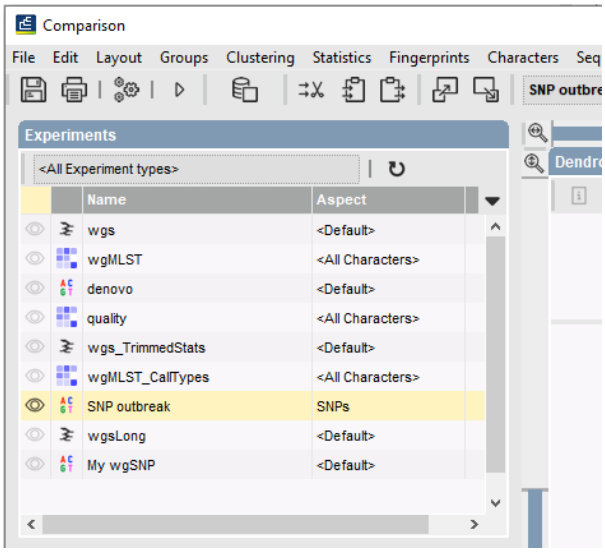

Figure 10: SNPs character aspect in the *Comparison* window.

We can now create a cluster analysis based on the SNP data, in the same way that a similarity-based clustering is performed in BIONUMERICS:

2. Make sure the **SNP outbreak** experiment is selected in the *Experiments* panel and select ***Clustering*** > ***Calculate*** > ***Cluster analysis (similarity matrix)...***.

Only multi-state similarity coefficients are suitable for clustering of SNP data. The ***Categorical (differences)*** coefficient treats each different value as a different state, and results in a distance matrix. With the ***Scaling factor*** one can deal with the hard-coded maximum of 200 that can be calculated for a distance value.

3. Select the ***Categorical (differences)*** coefficient, enter a ***Scaling factor*** of “100” and press <***Next***> (see Figure 11).

4. Select the ***Complete linkage*** clustering method and press <***Finish***> to calculate the dendrogram.

The resulting dendrogram is displayed in the *Dendrogram* panel of the *Comparison* window (see Figure 12).

5. To view the number of SNPs on the branches, select ***Clustering*** > ***Dendrogram display settings...*** (), and tick the option ***Show node information***. Press <***OK***>.

To trace back the number of SNPs from the branches or distance matrix, the displayed values needs to be multiplied with the ***Scaling factor*** used (here: 100).

6. To clear the selection, press the **F4**-key.

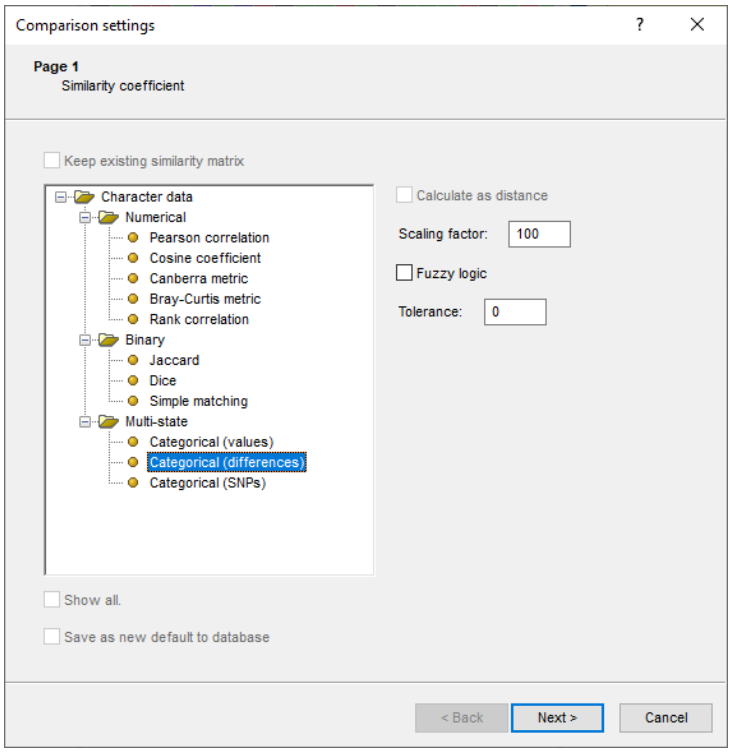

Figure 11: The categorical similarity coefficient.

The dendrogram contains one well-defined cluster and a number of unrelated strains.

7. Hold the **CTRL**-key and click on this cluster of related strains to select the 10 entries in the database.

This cluster contains all strains with MLST sequence type 22 (see info field '"ST info") and corresponds to the strains that were identified as part of the outbreak in the published study (see Figure 12).

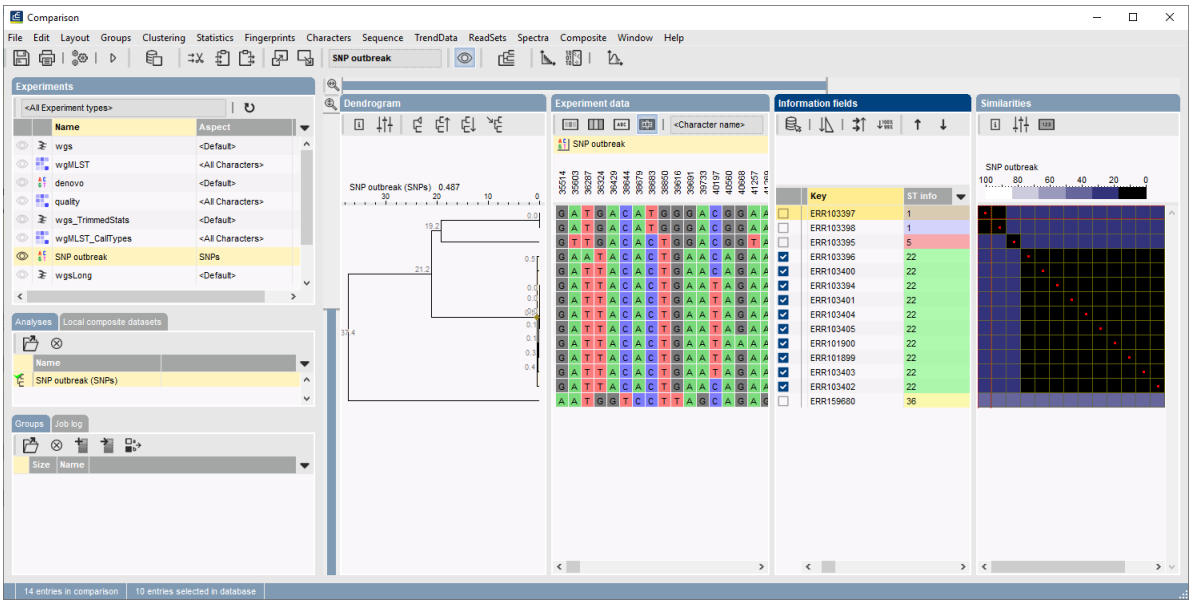

Figure 12: The *Comparison* window: wgSNP dendrogram.

8. Save the comparison with the dendrogram by selecting ***File*** > ***Save as...***. Specify a name (e.g. **Neonatal study**) and press <***OK***>.

4.2 Comparison with wgMLST results

For comparison purposes, we will create a cluster analysis for the same entries, but now based on wgMLST data:

9. Click on the icon left of **wgMLST** and select the ***wgMLST loci*** aspect from the drop-down list to visualize the wgMLST allele numbers in the *Experiment data* panel.

10. Select ***Clustering*** > ***Calculate*** > ***Cluster analysis (similarity matrix)...***, highlight the ***Categorical (values)*** coefficient and press <***Next***>.

11. Leave the default settings enabled and press <***Finish***> to calculate the dendrogram.

The resulting dendrogram is displayed in the *Dendrogram* panel of the *Comparison* window (see Figure 13).

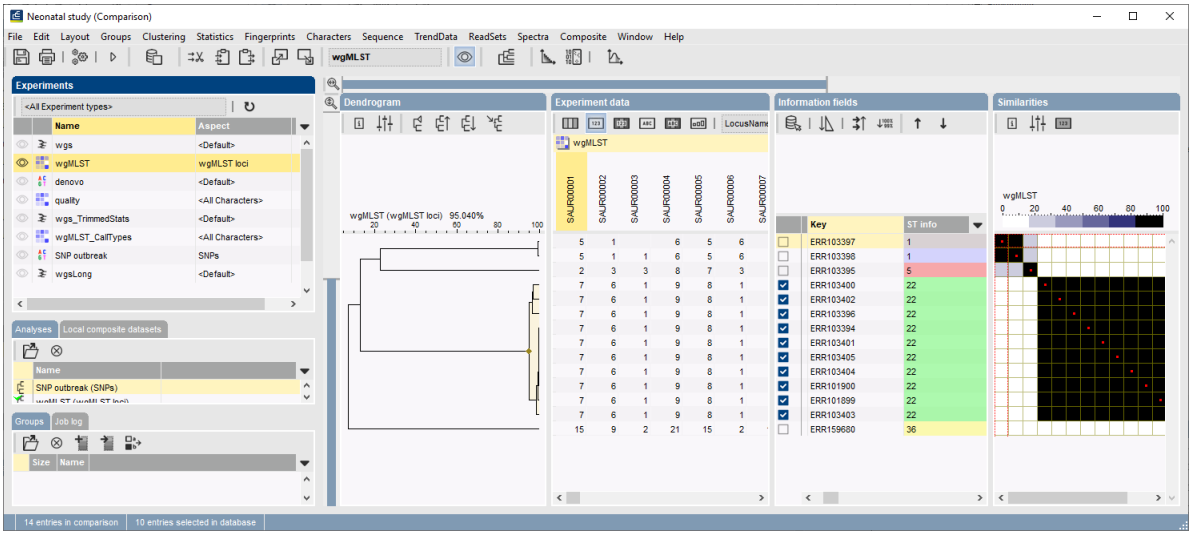

Figure 13: The *Comparison* window: wgMLST dendrogram.

Via the *Analyses* panel, you can switch back and forth between the two dendrograms. From this visual inspection, it is clear that both analyses correlate very well.

12. Select ***Clustering*** > ***Congruence of experiments...***.

Both experiments show a very high congruence of 95%.

13. Close the *Experiment congruence* window.

4.3 Zooming in on the outbreak

As an alternative to the procedure described previously, a SNP filtering can also be started from the *Comparison* window. We will illustrate this workflow by ”zooming in” on the isolates that belong to the outbreak:

14. First, make sure no entries are selected in the *Comparison* window by pressing **F4**.

15. Select the 10 entries that belong to the largest cluster in the dendrogram, e.g. using **Ctrl+click** on the corresponding branch in the dendrogram.

16. Return to the *Main* window to create a new comparison for the selected entries (highlight the *Comparisons* panel and select ***Edit*** > ***Create new object...*** (+)) or directly using the **Alt+C** keyboard shortcut.

A new *Comparison* window pops up with the ten entries that are associated with the outbreak.

17. Select ***File*** > ***Save*** (, **Ctrl+S**), enter e.g. **Outbreak** as name and press <***OK***>.

18. Click on the icon left of **SNP outbreak** to visualize the sequences in the *Experiment data* panel.

19. Select ***Sequence*** > ***Open SNP window...***

20. Highlight a SNP template from the *SNP analysis* wizard, e.g. ***Strict filtering*** (the same as we used previously).

This action shows the *SNP filtering* window again, as discussed previously.

21. Export the SNP matrix back to the comparison via ***File*** > ***Export to comparison...*** ( ).

This time, a ***SNP data name*** will be prompted for. Saving SNP matrices under different names will create multiple character aspects for the same sequence type. These aspects are available via the 'Aspect' drop-down list in the *Experiments* panel and can each be used to create cluster analyses from.

22. Enter e.g. **SNPs strict** and press <***OK***> in the dialog.

23. Return to the **Outbreak** comparison, where the SNP matrix is now displayed.

We can again calculate a dendrogram:

24. Select ***Clustering*** > ***Calculate*** > ***Cluster analysis (similarity matrix)...***.

25. Select the ***Categorical (differences)*** coefficient, specify a ***Scaling factor*** of "1" and press <***Next***>.

26. Check ***Complete linkage*** for ***Method*** and press <***Finish***> to calculate the dendrogram.

The resulting dendrogram indicates a sub-structure in the isolates that are associated with the outbreak: two well-defined groups are found (see Figure 14).

4.4 Exporting SNP data

If needed, SNP data can be exported as a character set. We will illustrate this for the **SNPs strict** aspect of **SNP outbreak**:

27. In the *Comparison* window, select **SNPs strict** from the 'Aspect' drop-down list next to **SNP outbreak**.

28. Select ***File*** > ***Export*** > ***Export character data...***.

29. In the *Export character data* dialog box, make sure ***Export mapped values*** is checked and press <***OK***>.

The exported SNP matrix will open automatically in MS Excel.

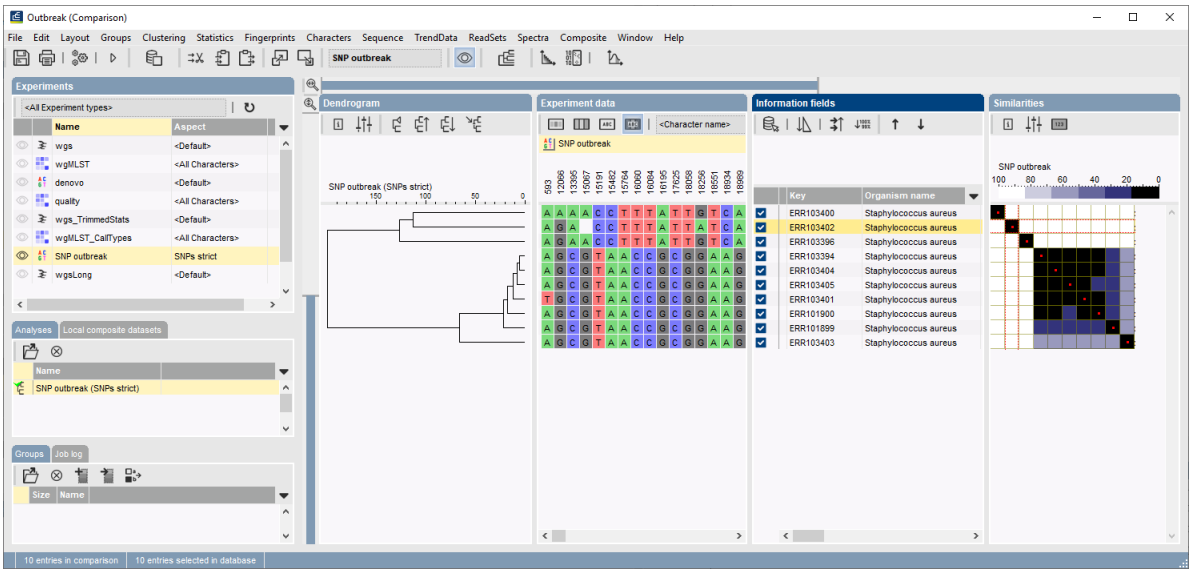

Figure 14: Zooming in on the outbreak: two groups.

Alternatively, the data displayed in the *SNP matrix* panel of the *SNP filtering* window can be exported using the column properties button (▾) and selecting e.g. ***Save content to file***.

Other applications might require the list of SNPs per entry formatted as (pseudo-)sequence:

30. In the *Comparison* window, with the **SNP Strict** aspect still selected, use ***File*** > ***Export*** > ***Export sequences nonredundant (fasta)***.

The `export.txt` file that opens is a multi-FASTA file with each row of the SNP matrix represented by a sequence.